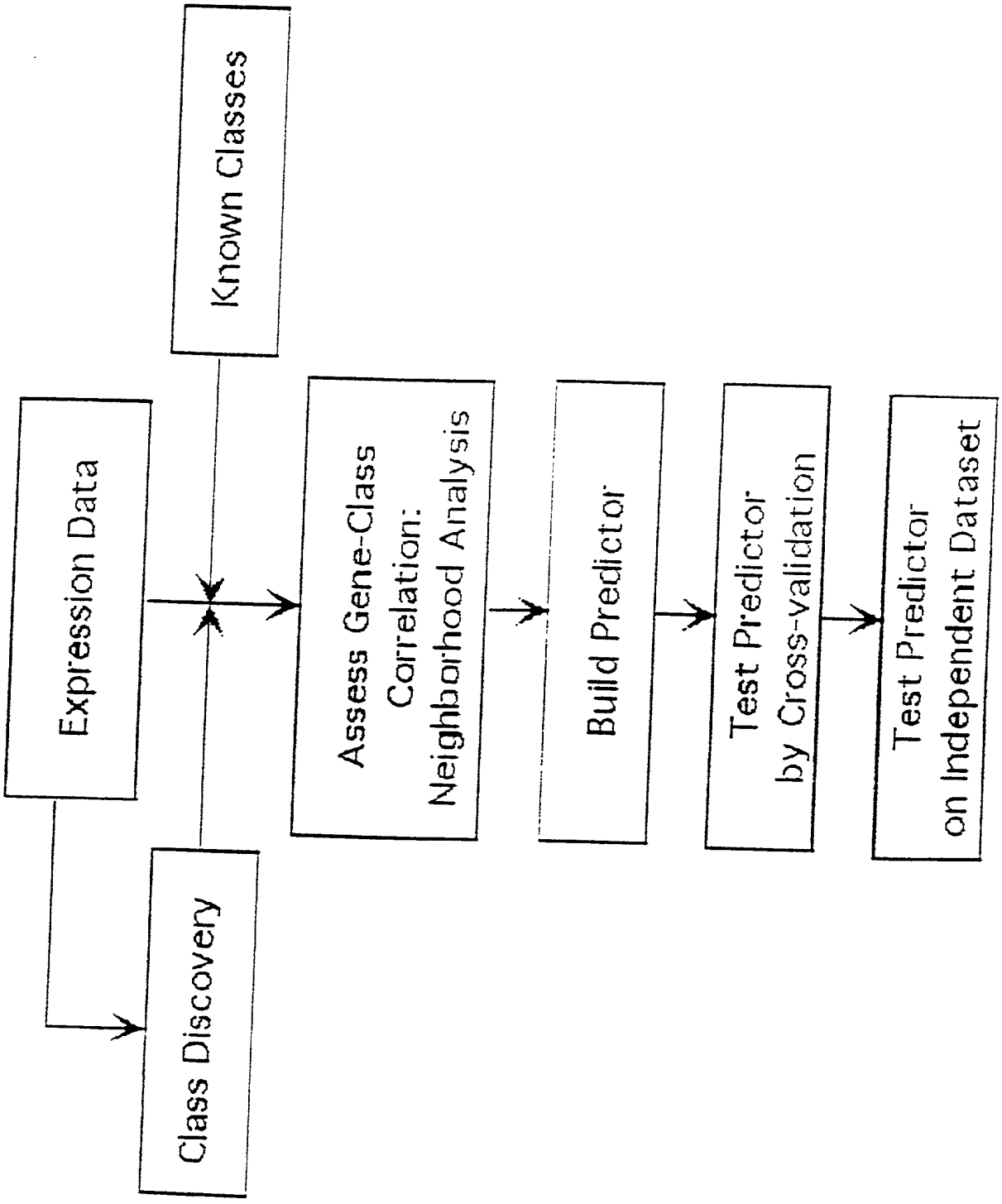

Known Classes
Expression Data
Assess Gene-Class Correlation: Neighborhood Analysis
Build Predictor
Test Predictor by Cross-validation
Test Predictor on Independent Dataset
Class Discovery

Figure 1A

Figure 1B

Weight
AML
ALL
V_{AML}
V_{ALL}
μ_{ALL}
μ_{AML}
gene1
gene2
gene3
gene4
gene5

Figure 1C

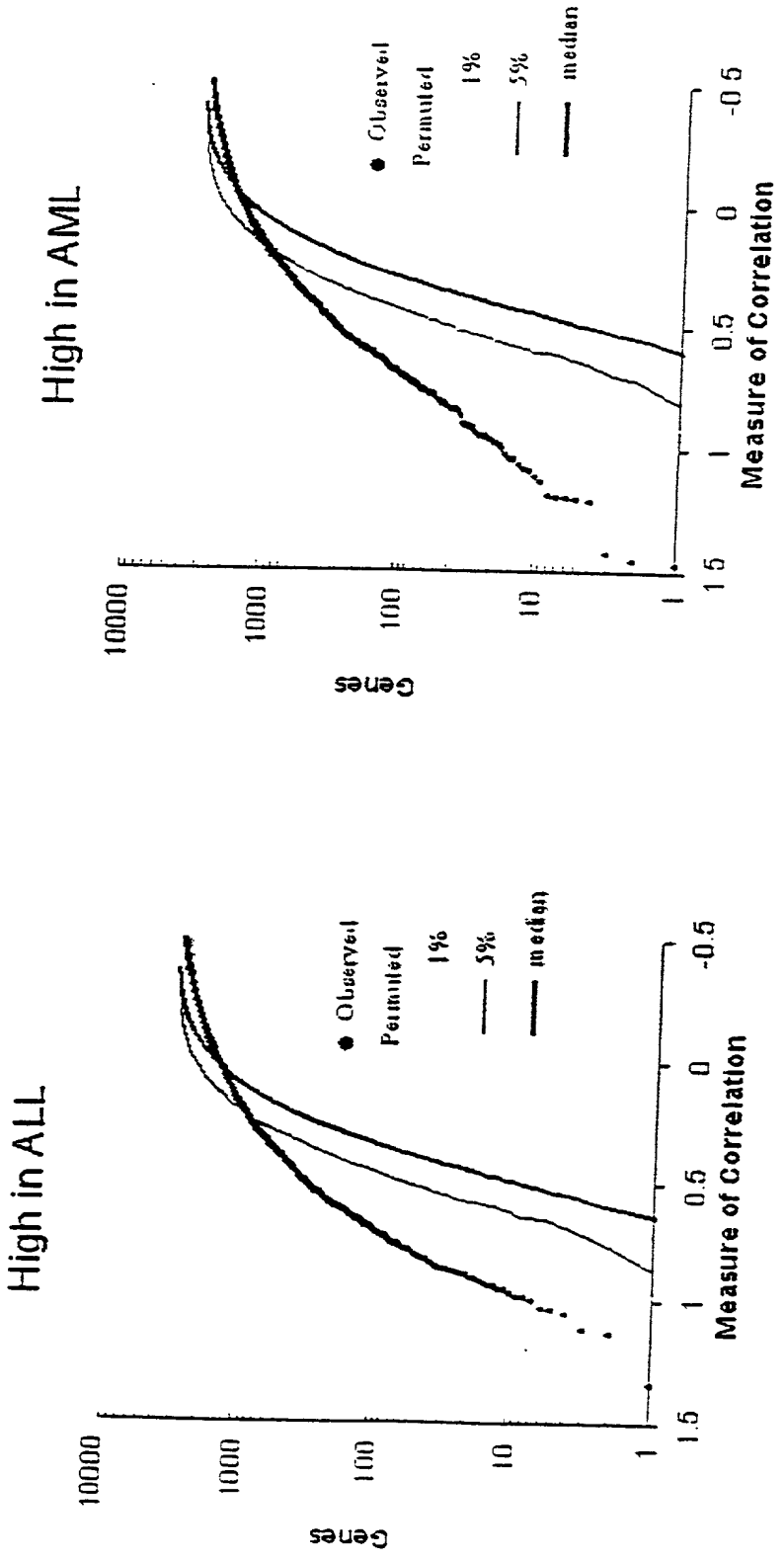

High in AML
High in ALL
Observed
Permuted
1%
5%
median
Genes
10000
1000
100
10
1
1.5
1
0.5
0
-0.5
Measure of Correlation

Figure 2

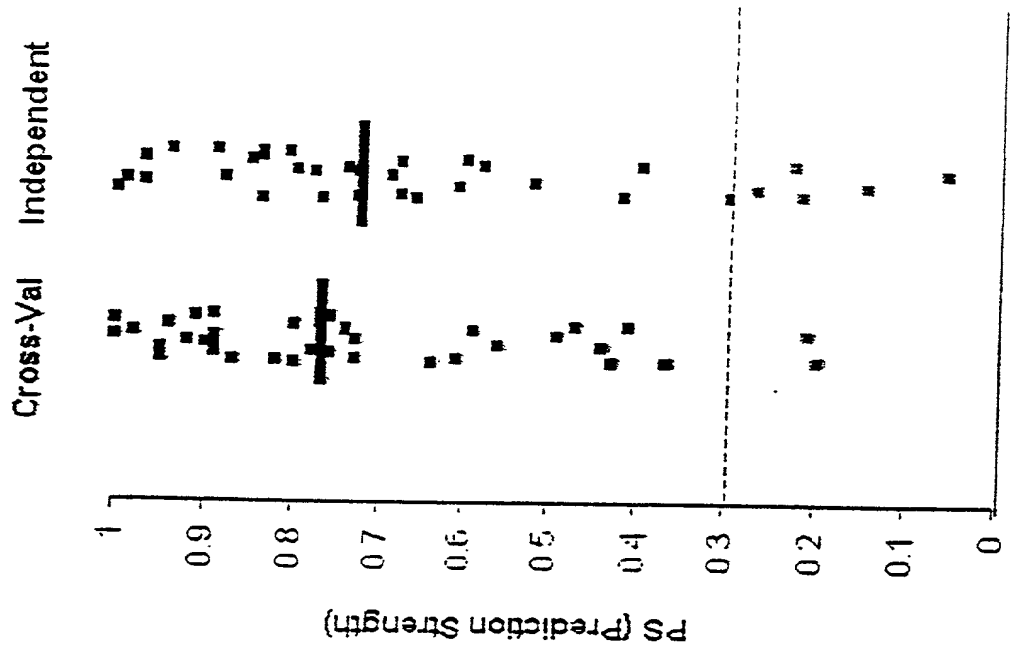

Cross-Val
Independent
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Fig. 3A

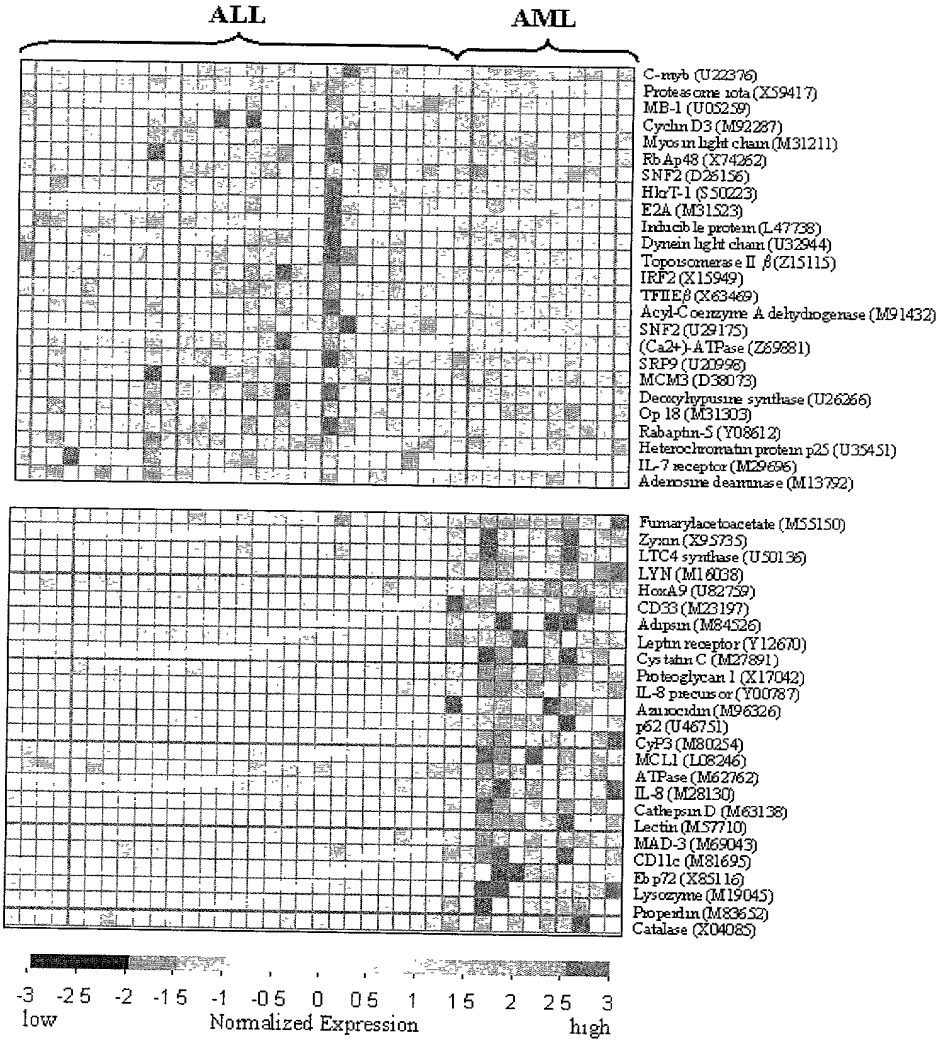
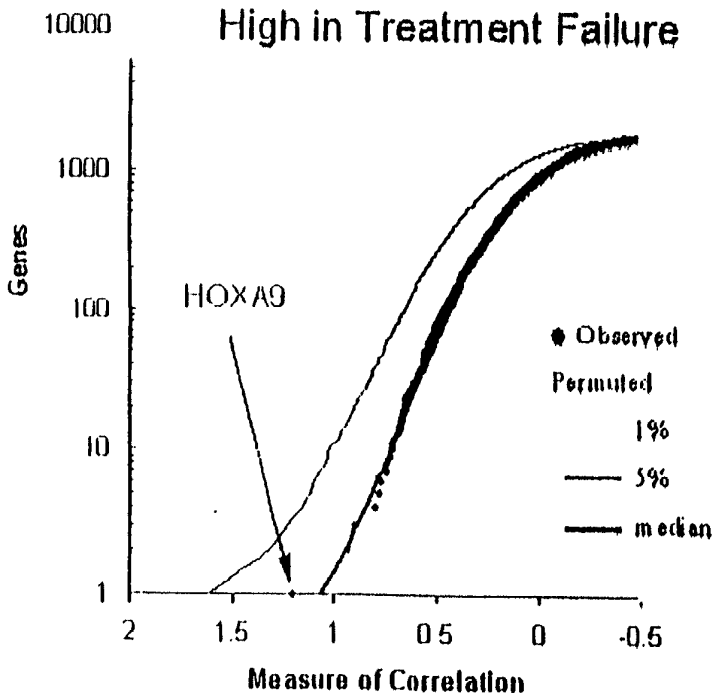

ALL
AML
C-myb (U22376)
Proteasome iota (X59417)
MB-1 (U05259)
Cyclin D3 (M92287)
Myosin light chain (M31211)
RbAp48 (X74262)
SNF2 (D26156)
HlaT-1 (S50223)
E2A (M31523)
Inducible protein (L47738)
Dynein light chain (U32944)
Topoisomerase II β (Z15115)
IRF2 (X15949)
TFIIEβ (X63469)
Acyl-Coenzyme A dehydrogenase (M91432)
SNF2 (U29175)
(Ca2+)-ATPase (Z69881)
SRP9 (U20998)
MCM3 (D38073)
Deoxyhypusine synthase (U26266)
Op 18 (M31303)
Rabaptin-5 (Y08612)
Heterochromatin protein p25 (U35451)
IL-7 receptor (M29696)
Adenosine deaminase (M13792)
Fumarylacetoacetate (M55150)
Zyxin (X95735)
LTC4 synthase (U50136)
LYN (M16038)
HoxA9 (U82759)
CD33 (M23197)
Adipsin (M84526)
Leptin receptor (Y12670)
Cystatin C (M27891)
Proteoglycan 1 (X17042)
IL-8 precursor (Y00787)
Azurocidin (M96326)
p62 (U46751)
CyP3 (M80254)
MCL1 (L08246)
ATPase (M62762)
IL-8 (M28130)
Cathepsin D (M63138)
Lectin (M57710)
MAD-3 (M69043)
CD11c (M81695)
Eb p72 (X85116)
Lysozyme (M19045)
Properdin (M83652)
Catalase (X04085)
-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
low
Normalized Expression
high

FIG. 3B


High in Treatment Failure
10000
1000
100
10
1
Genes
HOXA9
Observed
Permuted
1%
5%
median
2
1.5
1
0.5
0
-0.5
Measure of Correlation

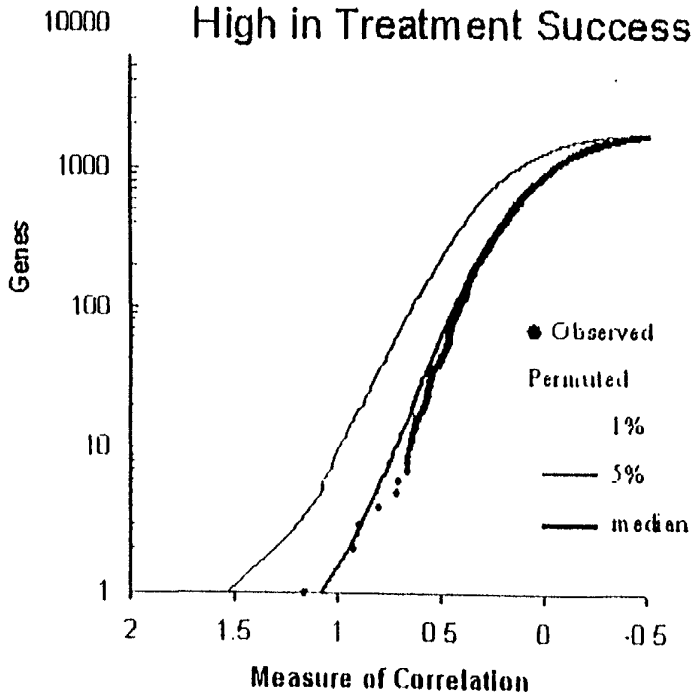

High in Treatment Success
10000
1000
100
10
1
Genes
Observed
Permuted
1%
5%
median
2
1.5
1
0.5
0
-0.5
Measure of Correlation

Figure 4

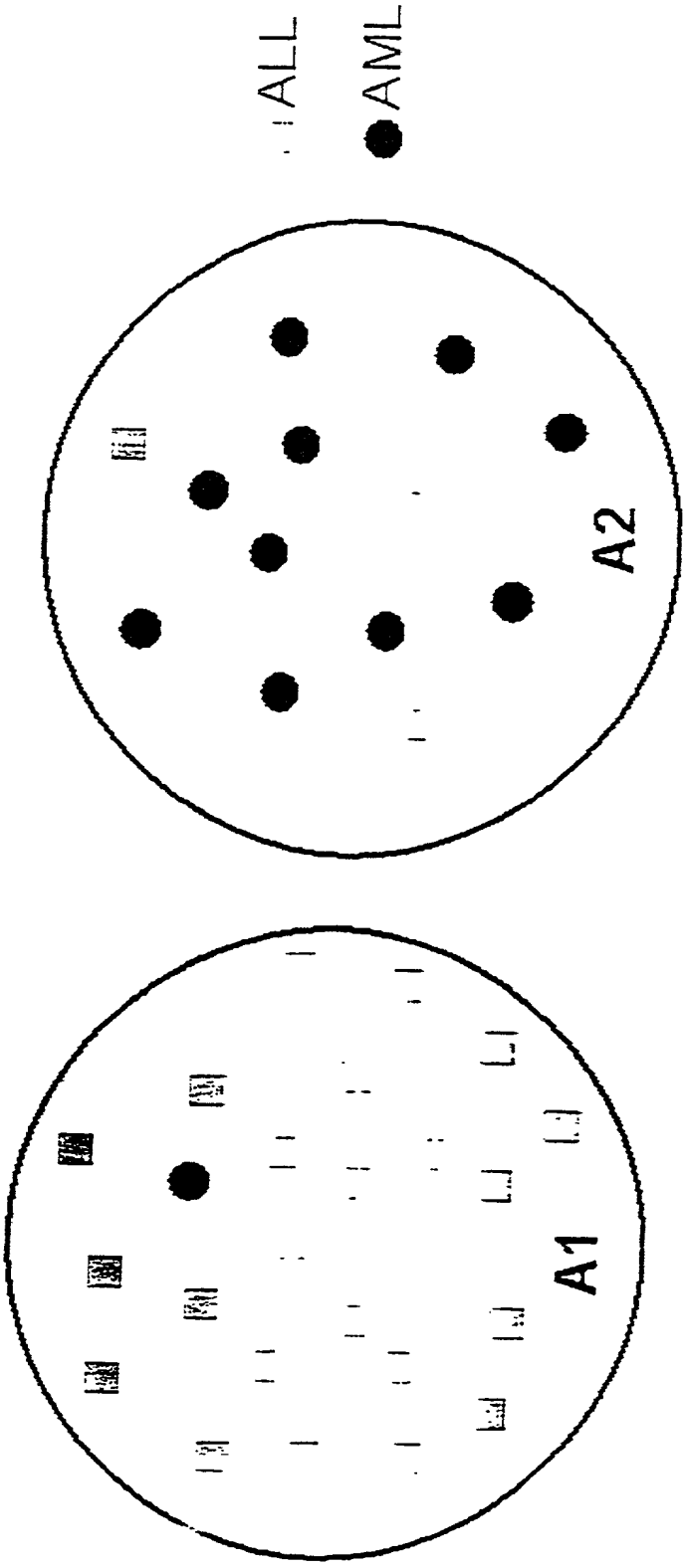


Figure 5A

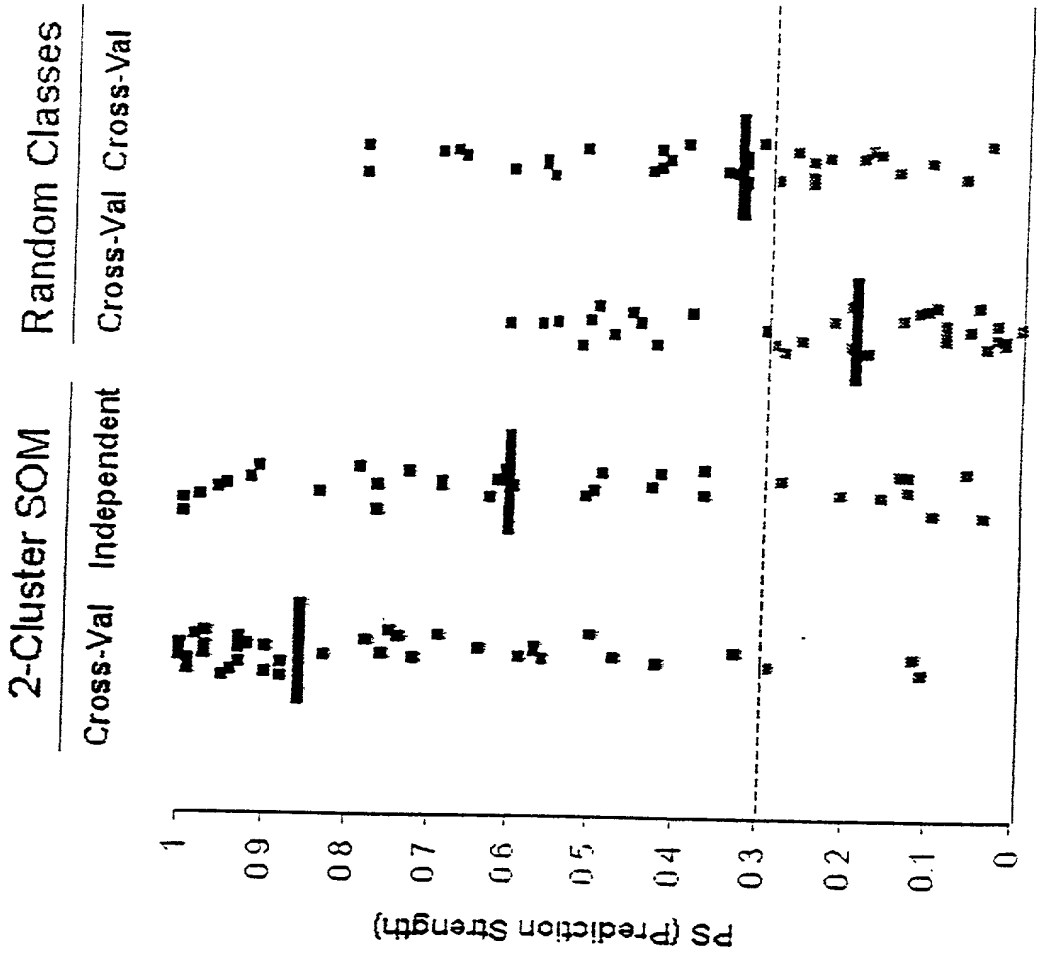

2-Cluster SOM
Random Classes
Cross-Val
Independent
Cross-Val
Cross-Val
PS (Prediction Strength)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure 5B

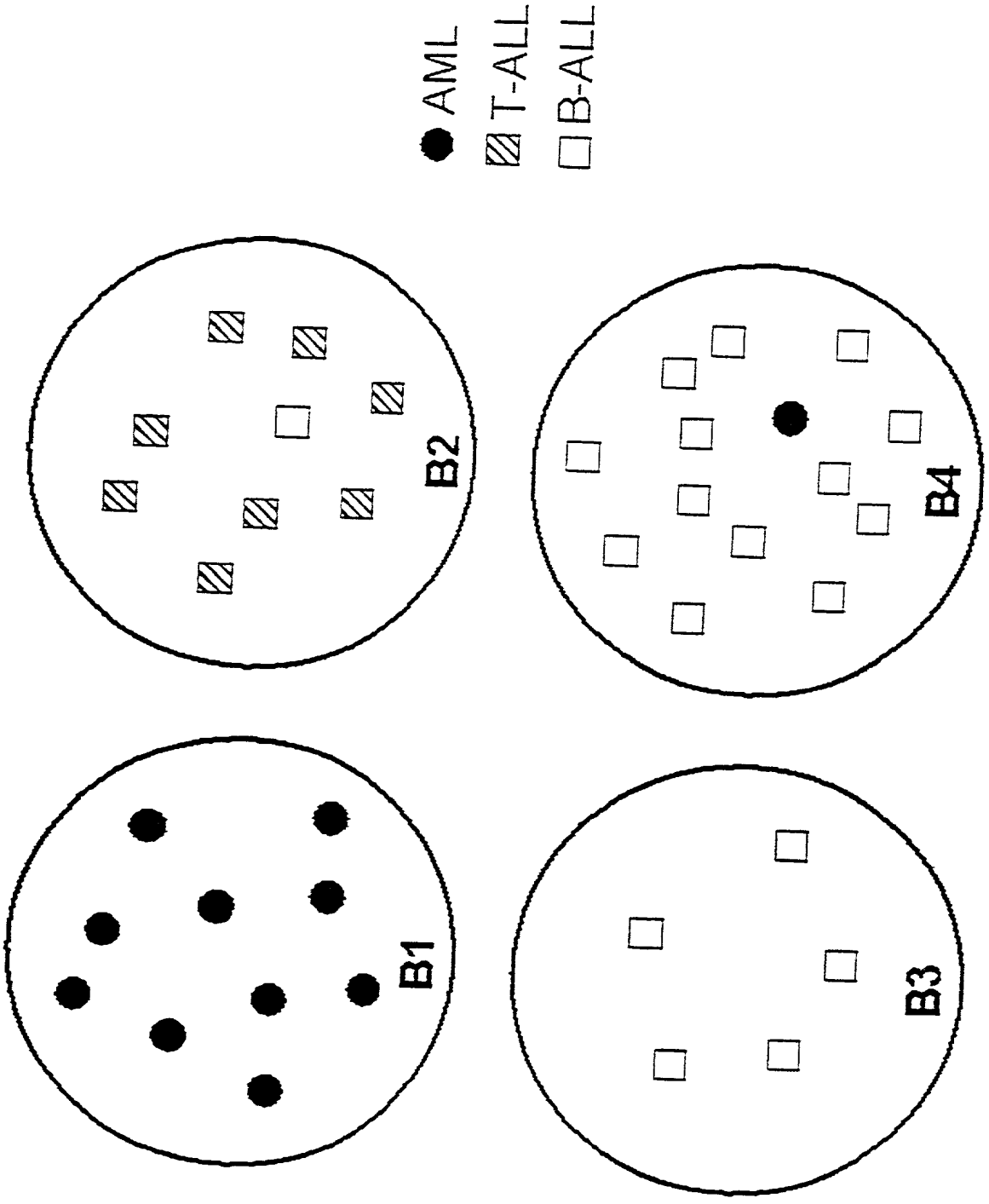


Figure 5C

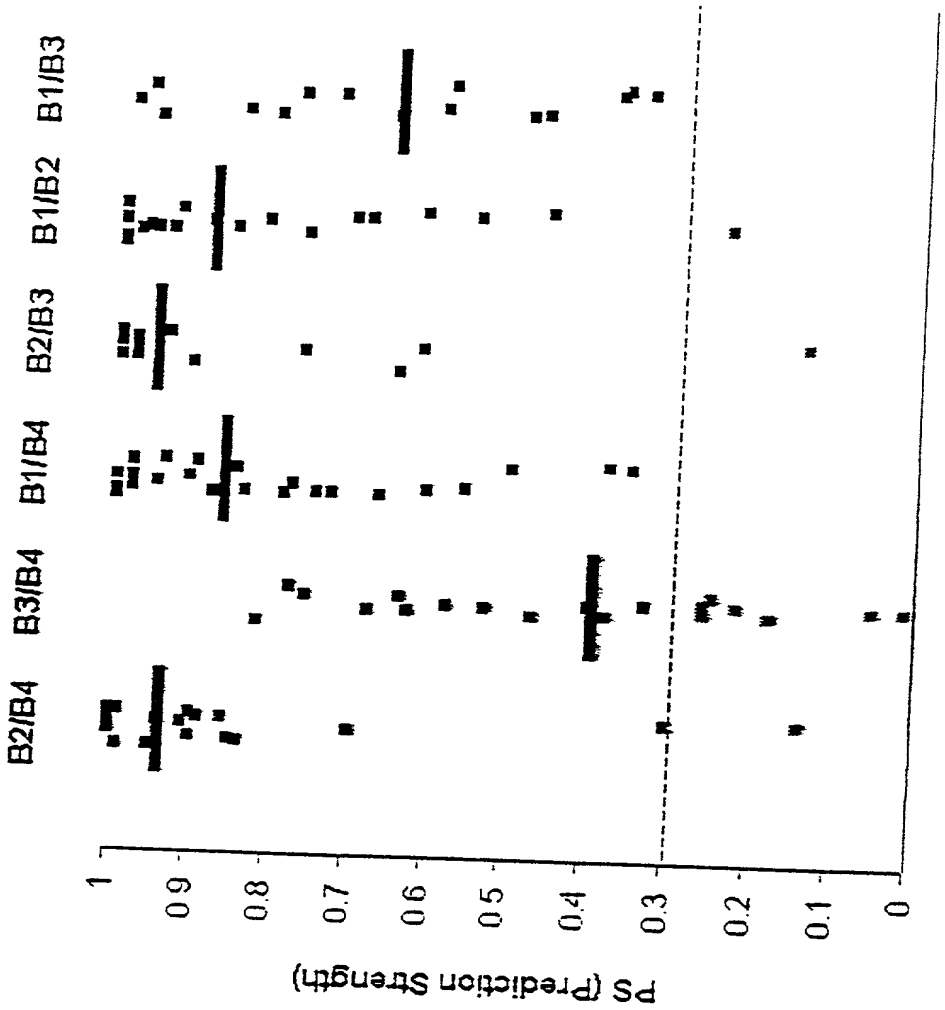

B2/B4
B3/B4
B1/B4
B2/B3
B1/B2
B1/B3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
PS (Prediction Strength)

Figure 5D

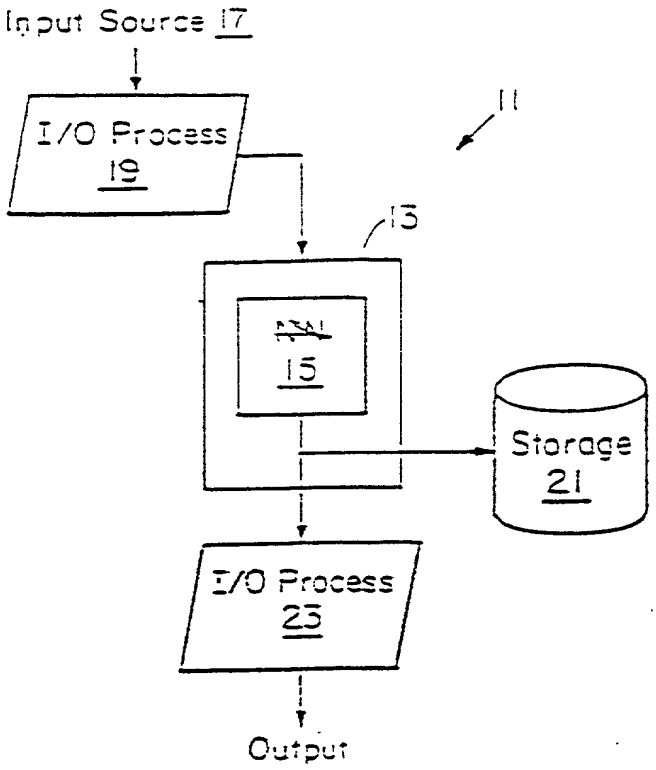

Input Source 17
I/O Process 19
11
13
15
Storage 21
I/O Process 23
Output

Figure 6

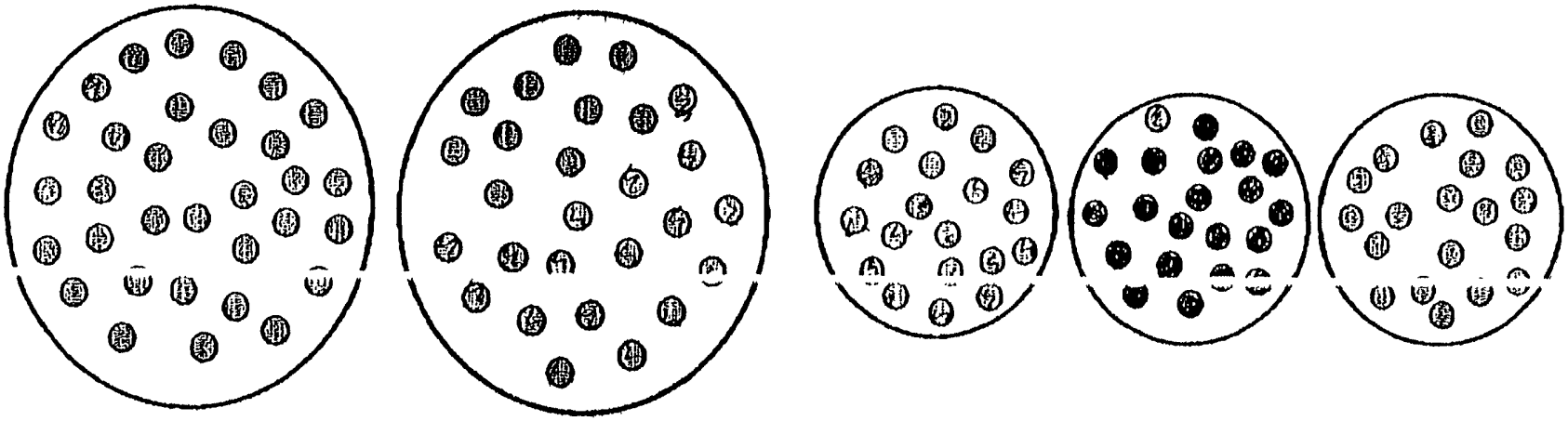
Example of Self-Organizing Map Class Discovery:

Lymphoma Large B-Cell and Follicular

Sample collection:

39 large B-Cell + 20 Follicular

Discovered Classes:



Large B-Cell

Follicular

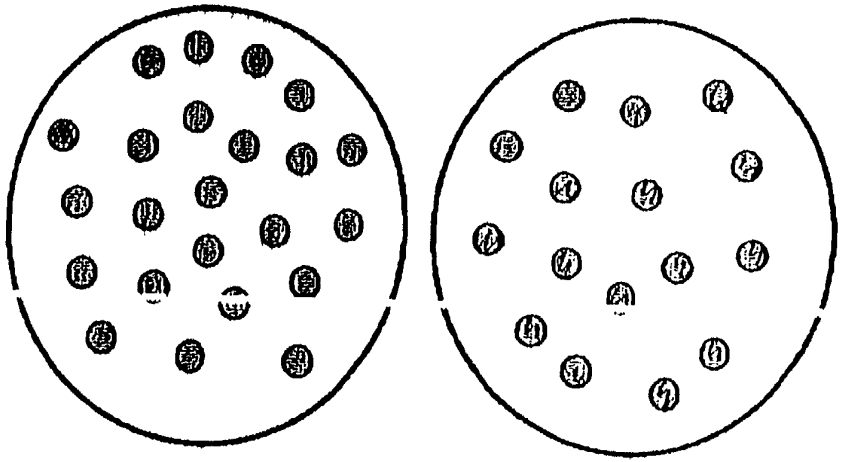
Fig. 7

Example of Self-Organizing Map Class Discovery: Brain Glioma and Medulloblastoma

Sample collection:

24 Medulloblastomas + 15 Gliomas

Discovered Classes:



Medulloblastoma

Glioma

Fig. 8

Multidimensional Scaling of Leukemia Samples

(431 genes, 15-fold, delta > 1500, thres=100, ceil=16,000)

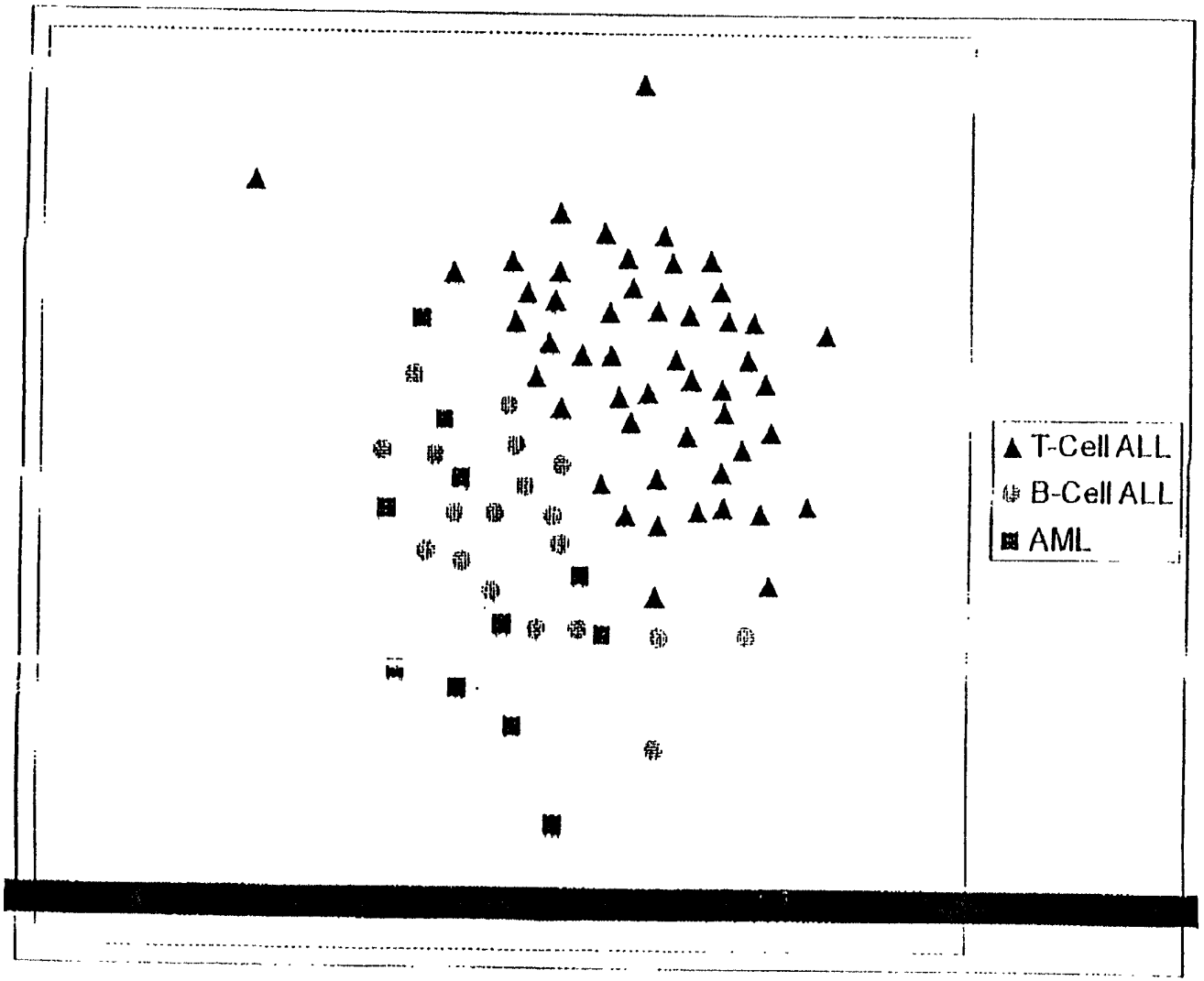


Fig. 9

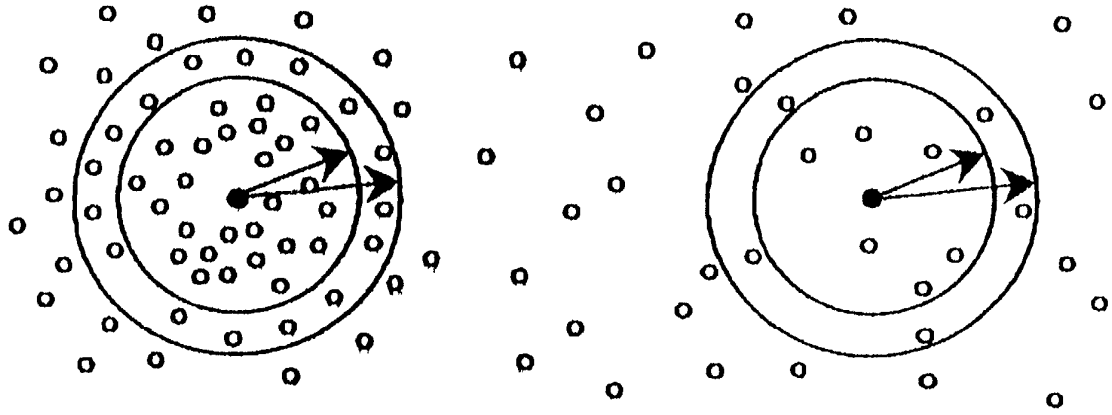
Hierarchy of Problems in Molecular Class Prediction

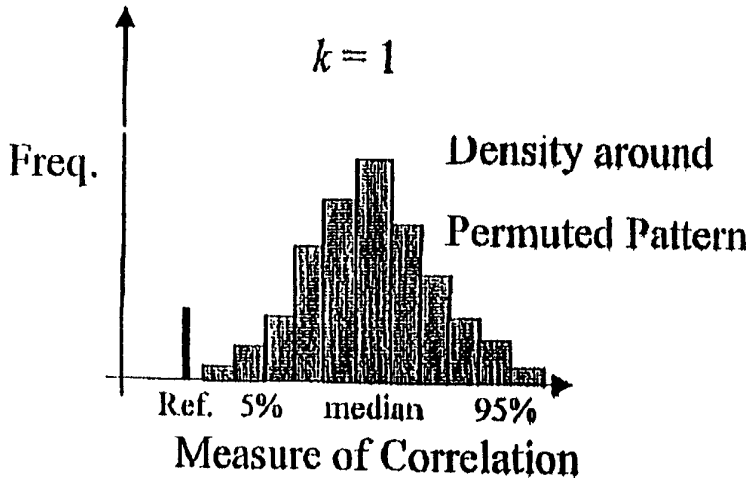
Problem:	Difficulty:	Gene Markers:	Error:	Example:
I. Tissue or Cell Type Normal vs. Abnormal	Low	~1000-2000	~0%	Normal vs. Renal Carcinoma 
II. Morphological Type	Low-medium	~200-500	~0-5%	Leukemia ALL vs. AML 
III. Morphological Subtype	Medium-high	~50-100	~0-15%	ALL B- vs. T-Cell 
IV. Treatment Outcome Drug Sensitivity	High	~1-20	~5-50%	AML Treatment Outcome 

Fig. 10

Docket No.: 2825.1018-010
Title: "Methods for Classifying ..."
Inventors: Todd R. Golub, *et al.*

Neighborhood Analysis: Assessing Statistical Significance of Gene-Class Correlations





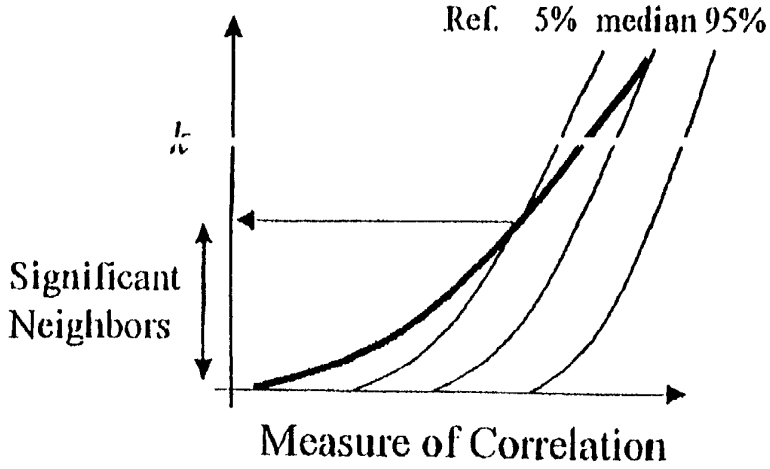


Fig. 11

Class Prediction Results

Problem Type	Biological System	Problem Distinction	Number of samples	Number of errors	Number of no calls	Number of errors (all calls)	Number of gene markers
	Renal	Normal vs Cadroma	12	0 (0%)	0	0 (0%)	>100
	Leukemia	ALL vs AML	35	0 (0%)	2	0 (0%)	200
	Leukemia	ALL B vs T Cell	38	0 (0%)	1	1 (3%)	200
	Leukemia	Treatment Outcome	15	2 (13%)	0	2 (13%)	1

Fig. 12